 doi:10.1093/bib/bbu005

Editorial: Alignment-free methods in computational biology

Alignment-free methods for biological sequence analysis and comparison have emerged as a natural framework to address the challenges of understanding the patterns and properties of biological sequences. These methods are based on mapping symbolic sequences describing DNA, RNA and proteins, onto vector spaces, in which many of the analysis can be performed more efficiently. The rational is to represent sequences as numerical real-valued vectors and to apply available tools to this domain, which range from filtering techniques, normalization, dissimilarity estimation and clustering. Broad frameworks such as probability, statistics and linear algebra are then at hand to provide a strong and extensive theoretical and computational background. This explains the high computational efficiency of alignment-free methods, which has led in the past decades to widening the range of successful applications.

The main advantage of alignment-free methods, besides the key fact they are usually computational inexpensive, is the ability of effortlessly dealing with whole genomes, thus allowing the analysis of complete sequence information. They are robust to shuffling and recombination events and generally applicable when less conservation pushes beyond what alignment could handle. It should nevertheless be noted that for the majority of the alignment-free algorithms, the symbol order is lost, which might constitute a caveat if an alignment structure is preferred.

The term alignment-free was coined in a review paper a decade ago [1] that reflected on research being performed by then, which avoided the caveats of dynamic programming. That survey tried to systematize, organize and categorize a diversity of disparate methods in a common framework. It was then observed that all the main categories identified shared, in their root, the same principle and rationale of not preprocessing the sequences being compared by aligning them. It was already clear that there was a central focus on vector-valued representations using L-tuple composition, sequence representation through iterative maps, compression and entropy estimation, although with contrasting notations and generally unaware of the possibility of a unifying perspective. The applications surveyed were surprisingly diverse, ranging from phylogenetic classification and motif analysis to genomic entropy estimation.

This special issue on *Alignment-free methods in computational biology* reflects this growing trend and offers current reviews on several areas where alignment-free methods continue to provide relevant results in computational biology and bioinformatics. The papers included cover a wide area of this theme and constitute a tentative roadmap for future developments, which both go far beyond biological sequence analysis, and also are more in line with the new Big Data challenges brought about by next-generation sequencing.

The statistical aspects of metrics designed for sequence comparison are described in *New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing* by Song, Ren, Reinert, Deng, Waterman and Sun. Previous definitions of adequate dissimilarly measures and metrics to apply in comparison tasks led to the development of statistical descriptions and strong theoretical work, addressed in this article. The authors review statistical properties of metrics based on L-tuple matches, in particular D2 statistic, illustrating its strength for clustering purposes of next-generation sequencing short read data in metagenomic studies.

The close relation between alignment-free methods and pattern recognition algorithms is reviewed in *Pattern recognition and probabilistic measures in alignment-free sequence analysis* by Schwende and Pham, where the authors survey the general properties of alignment-free versus alignment-based methods and overview current metrics and software to perform the general analysis in the context of machine learning.

The need for alternative sequence representation has led to the striking development of iterated maps, rooted in non-linear dynamics, reviewed in the paper *Sequence analysis by iterated maps, a review* by

© The Author 2014. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/15/3/341/186352 by guest on 17 April 2024

Almeida. The idea of mapping sequences onto vector spaces attained a high level of formal elegance in chaos game representation. These functions allowed to smoothly bridge concepts and applications from numerical representation to graphical structures. Iterated maps have also clear connections with stochastic processes and Markov chain models, besides owning strong links with information theory concepts. The developments in the past decade suggest a particularly intriguing angle on the scalable analysis of next-generation sequencing.

The alignment-free methods covered have been particularly connected to information theory (IT) concepts. The statistical and linear algebra descriptions have clear associations with IT and both have been successfully applied in computational biology. The paper *Information theory applications for biological sequence analysis* by Vinga reviews and categorizes IT applications under an alignment-free framework. These include the global characterization of sequences, their local analysis and also methods that combine several levels of information into a unique integrated framework.

In connection with information theory concepts, the algorithmic assessment of compression methods for sequences is addressed in the paper *Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies* by Giancarlo, Rombo and Utro. The authors exhaustively review current compression and storage techniques, with a focus on high-throughput sequencing (HTS) data. They further provide reference databases and available software tools.

On the application to evolutionary research, the paper *Alignment-free phylogenetics and population genetics* by Haubold reviews alignment-free methods successfully applied to phylogenetics and population genetics. The author overviews the metrics more adequate to infer phylogenetic relationships and to estimate the distribution of mutations, and illustrates them in simulated sequences and in real genomes.

Finally, the paper *Applications of alignment-free methods in epigenomics* by Pinello, Lo Bosco and Yuan illustrates the strength of this framework in the context of linking genome to epigenome. Several machine learning techniques are overviewed and their applications highlighted, namely for nucleosome positioning, DNA methylation and histone modifications.

We hope that this special issue with current reviews of alignment-free methods will support algorithm advancements in the next decade as exciting as in the 11 years since the establishment of a unifying characterization.

Susana Vinga
IDMEC, Instituto Superior Técnico - Universidade de Lisboa (IST-UL) Av. Rovisco Pais, 1049-001 Lisboa, Portugal
E-mail: svinga@dem.ist.utl.pt

Reference

1. Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics* 2003;**19**:513–23.

Downloaded from https://academic.oup.com/bib/article/15/3/341/186352 by guest on 17 April 2024